

# Inductive Logic Boosting

Wang-Zhou Dai, Zhi-Hua Zhou\*

*National Key Laboratory for Novel Software Technology  
Nanjing University, Nanjing 210093, China*

---

## Abstract

Recent years have seen a surge of interest in Probabilistic Logic Programming (PLP) and Statistical Relational Learning (SRL) models that combine logic with probabilities. Structure learning of these systems is an intersection area of Inductive Logic Programming (ILP) and statistical learning (SL). However, ILP cannot deal with probabilities, SL cannot model relational hypothesis. The biggest challenge of integrating these two machine learning frameworks is how to estimate the probability of a logic clause only from the observation of grounded logic atoms. Many current methods model a joint probability by representing clause as graphical model and literals as vertices in it. This model is still too complicate and only can be approximate by pseudo-likelihood. We propose Inductive Logic Boosting framework to transform the relational dataset into a feature-based dataset, induces logic rules by boosting Problog Rule Trees and relaxes the independence constraint of pseudo-likelihood. Experimental evaluation on benchmark datasets demonstrates that the AUC-PR and AUC-ROC value of ILP learned rules are higher than current state-of-the-art SRL methods.

---

## 1. Introduction

In recent years, there has been an increasing interest in integrating first-order logic with probabilities by defining confidence of a logic formula with a weight. This interest has resulted in the fields of Probabilistic Logic Programming (PLP) [DRK08] and Statistical Relational Learning (SRL) [GT07]. PLP focuses on the extension of logic programming languages with probabilities,

---

\*Corresponding author. Email: zhouzh@nju.edu.cn

as in for instance Problog [DRKT07]. Conversely, SRL extends first-order logic on probabilistic graphical models as Markov or Bayesian networks, like Markov logic Network (MLN) [RD06].

Structure learning of PLP and SRL is an important but challenging task because it defines the complex relationships among entities and have to be learned in a exponential searching space. The task actually is a intersection of Inductive Logic Programming (ILP) and Statistical learning (SL).

Inductive Logic Programming [Mug91] is a combination of inductive learning and logic programming, which employs techniques from both machine learning and logic programming. Given a set of training examples of a target predicate(relation) and background knowledge, ILP finds a hypothesis which is complete and consistent with evidence and background knowledge[LD94].

ILP can learns relational data which is a collection of logic relations, while SL deals with datasets of grouped instances that have distinguishing feature values. The difference between these two tasks makes the combination of ILP and SL very difficult. However, structure learning of PLP and SRL is so important that has received much attention recently, many practical algorithms have been proposed.

Most of these methods assume a joint distribution on a probabilistic graph model for each logic formula, then learn parameters that maximizing pseudo-likelihood of the formulas by exploiting the labeled examples. The learning strategy of these approaches can be categorized into two kinds. The first category learns logic formulas and weight separately, for instance as [KD05, MM07, KD09, KD10]. The other category turns the problem into a series of relational regression problems and learns the weights and the clauses simultaneously [KNKS11, NKK<sup>+</sup>12].

Graphical model based methods introduce pseudo-likelihood by assuming independence between logical literals, then turns the structure learning problem into a statistical learning problem and solve it with many statistical algorithms. However, this kind of transformation does not exploit the ability of induction of first-order logic. It treats logic literals independently during learning thus loses accuracy, it also sacrifices the expressiveness of logic because only uses statistical model to represent the relational hypothesis.

In this paper, we present Inductive Logic Boosting (ILB), possibly a new class of approaches to integrate ILP with statistical learning. It first finds paths for positive examples in a hyper-graph

that constructed from relational database. Then substitutes the paths to first-order core forms (patterns) to generate binary labeled feature-based instances. Finally learns an ensemble model through Adaboost the Problog Rule Trees for target predicate on the generated training data with specific features. The generated dataset and boosting introduces characters of statistical learning, Problog Rule Tree preserves features of ILP. ILB relaxes the assumption of independence and learns an PLP model in global underlying data distribution. Experiments shows ILB produces PLPs more accurate and more comprehensible than current state-of-the-art approaches.

The remainder of this paper is arranged as follows. We begin by reviewing related works in 2, then introduce some some backgrounds in 3. We describe the detail of ILB in 4 and report the experiments in 5. Finally, we conclude with future works in 6.

## 2. Related Work

The recent years of ILP have been dominated by the development of methods for learning probabilistic logic representations. A general framework for Probabilistic ILP (PILP) was introduced [DRK08].

Most of current systems integrates ILP and statistical learning by expressing first-order logic as probabilistic graphic models and then learn the parameters on the graph models. They search structures (candidate clauses) first, then learns the parameters (weights) and modify the structures (clauses) accordingly. This kind of approaches performs either top-down [KD05] or bottom-up searches [MM07]. There are also works learns PLP by beam search or approximate search in the space of probabilistic clauses [BR13, DBR13].

There are also some methods combines ILP with SL by boosting. For example, Boosting FFOIL [Qui96] directly adopts the boosting framework with a classical ILP system, FFOIL, as weak learners, it proves that boosting is beneficial for first-order induction. More recently, RDN-Boost [NKK<sup>+</sup>12] and MLN-Boost [KNKS11] turns the problem into relational regression problems and learns both structures and weights of graphical model simultaneously.

Different with previous methods, Inductive Logic Boosting transforms relational dataset to a feature-based dataset, then learns Problog Rule Trees by discriminative learning to induce both

first-order logic rules and their weights, finally use Adaboost to get an accurate hypothesis definition of target predicates.

### 3. Preliminary

In first-order logic, *formulas* are constructed by four types of symbols: *constants*, *variables*, *functors* and *predicates*. In this paper we follow the Prolog terminology that using words begin with a lowercase letter to represent *constants* and *predicates*, words begin with a uppercase letter to represent *variables*. A *term* is a variable, a constant, or a functor applied to terms. An *atom* is of the form  $p(t_1, \dots, t_n)$  where  $p$  is a predicate of arity  $n$  and the  $t_i$  are terms. A *formula* is built out of atoms using using quantifiers  $\forall, \exists$  and usual logical connectives  $\neg, \wedge, \vee, \rightarrow$  and  $\leftrightarrow$ . A *rule* (also called a *normal clause*) is a universally quantified formula of the form  $h : \neg t_1, \dots, t_n$ , where atom  $h$  is called the *head* of the rule and literals  $t_1, \dots, t_n$  the *body* where  $t_i$  are logical atoms. The formula means the conjunction  $t_1 \wedge \dots \wedge t_n$  will deduce  $h$ . A *logic program (LP)* is a set of FOL rules. A *fact* is a *rule* with an empty *body* and is written more compactly as  $h$ , means  $h$  is always true in the program.

The task of inductive logic boosting is similar with ILP, which can be formally put as this: Given (i) a set of training examples  $\mathcal{E}$ , including *true* groundings  $\varepsilon^+$  and *false* groundings  $\varepsilon^-$  of a target predicate(relation)  $p$ ; (ii) a description language  $\mathcal{L}$ , specifying syntactic restrictions on the definition of predicate on the definition of predicate  $p$ ; (iii) background knowledge  $\mathcal{B}$ , defining other predicates  $q_i$  that may be used in definition of  $p$ . Find a hypothesis  $\mathcal{H}$  as the definition of  $p$ , which can predict the confidence of each grounding of  $p$ .

Problog is one of Probabilistic Logic Programming languages. It integrates logic program with probability by adding a probability  $p$  to each ground facts  $f$ , written  $p :: f$ . It also allows *intentional* probabilistic statements of the form  $p :: p(A_1, A_2, \dots, A_n) : \text{-body}$ , where  $p(\cdot)$  is a probabilistic atom as head, *body* is a conjunction of calls to non-probabilistic facts. Like Prolog, the rules are range-restricted: all variables in the head of a rule should also appear in a positive literal in the body of the rule.

A Problog program specifies a probability distribution over Herbrand interpretation, or *Possible World*. The ground probabilistic fact  $p :: f$  gives an *atomic choice*, it means the program can

choose to include  $f$  as a fact with probability  $p$  or reject it with probability  $1 - p$ . A *total choice* is obtained by making an atomic choice for each ground probabilistic fact. The probability distribution over the total choices is defined to be the product of the probabilities of the atomic choices that it is composed of as independent events, i.e. there are two probabilistic facts  $0.3 :: a$  and  $0.8 :: b$ , then the total choices are  $\{a, b\}, \{a\}, \{b\}$  and  $\{\}$  with probabilities 0.24, 0.06, 0.56 and 0.14.

## 4. Proposed Approach

We now present the Inductive Logic Boosting framework for integrating logic induction and statistical learning. ILB represents Problog rules as decision trees, which we called *Problog Rule Tree*. It learns rules and weights simultaneously from an alternated dataset, which is a collection of a binary labeled instances generated from original relational data. Finally boosts on these rule sets adaptively to provide an accurate hypothesis.

### 4.1. Problog Rule Tree

A *Problog Rule Tree*  $T = (h, N, C)$  is a kind of Relational Probability Tree (RPT) [NJFH03].  $h = p(A_1, \dots, A_n)$  is the *head* of  $T$ , represents the target predicate  $p(A_1, \dots, A_n)$  to be earned,  $\{A_i\}$  are the arguments of  $p$ ;  $N = n_1, \dots, n_l$  is the node set of  $T$ . Each node  $n$  is either a decision node or end node. Each decision node  $n^d = \{t_1, \dots, t_k\}$  is a conjunction of logic atoms, it has a *true child* and a *false child* to determine whether an instance satisfies  $n^d$ , end node  $n^e$  records the proportion of positive instances which satisfy all its true ancestors. If  $n'$  is the true child of  $n$ , then  $n$  is called *true parent* of  $n'$ ;  $C = \{t_1, \dots, t_m\}$  is the tree root, together with  $h$  they formulate a short logic rule  $h: -C$  which we call *core form* of  $T$ .  $C$  is the first decision node of a Problog tree and only has a true child. Therefore, one Problog Rule Tree is learned to expand only one core form. In order to make the learned rule legal in Problog, we constrain that all variables  $A_i$  appear in  $h$  must also appear in  $C$ . For example, figure 1 denoted a tree to expand *core form*  $\text{sametitle}:-\text{hasword}(X, Z1), \text{hasword}(Y, Z1)$ .

From each end node in  $T$ , we can restore a Problog statement by backtracking its route to the root. For example in figure 1, we can get a statement from the 0.147 end node:

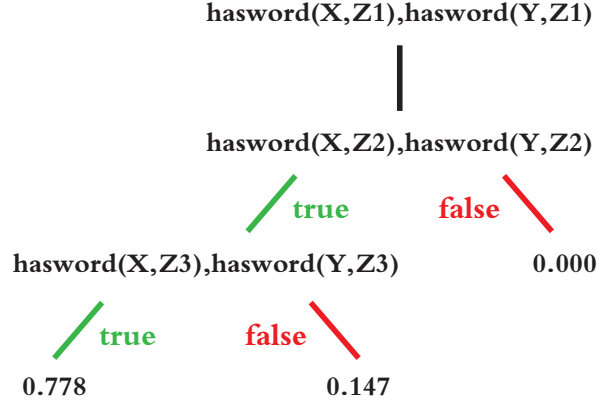


Figure 1: An Example of learned Problog Rule Tree for predicate  $h=\text{sametitle}(X, Y)$ : the root conjunction is the *core form*  $c$  that to be expanded, which means  $X$  and  $Y$  have a common word  $Z_1$ . Other clauses are decision nodes. The end nodes 0.000, 0.147, 0.778 correspond to the proportion of true  $\text{sametitle}(X, Y)$  that have only one, two or three common words, the proportions also are the probability of corresponding Problog rules.

0.147:: $\text{sametitle}(X, Y) :-$   
 $(\text{hasword}(X, Z_1), \text{hasword}(Y, Z_1)),$   
 $(\text{hasword}(X, Z_2), \text{hasword}(Y, Z_2)),$   
 $\backslash + ((\text{hasword}(X, Z_3), \text{hasword}(X, Z_3))),$   
 $\text{unique}([Z_1, Z_2, Z_3]).$

which means if  $X$  and  $Y$  only have 2 common words, the probability of  $\text{sametitle}(X, Y)$  is 0.147 (This rule is only for demonstration, the actual problog rule used in ILB will ignore negations of conjunctions).

The weighted proportion of positive instances in an end node is used as confidence of the statement. This is because for an individual tree, Prolog rule extracted from each end node covers different part of data (which is the nature of decision tree). Without overlap in the underlying distribution, the proportion of positive instances is the maximum likelihood estimation for the confidence of those rules.

Remind that each tree only deals with the formulas learned from one core form, although the expanded Problog rules from same core form has no intersection in their coverage, the trees who expand different core forms may cover same examples. For instance in entity resolution task, an example  $\text{sameauthor}(\text{person1}, \text{person2})$  may be covered by rule “0.2:: $\text{sameauthor}(X, Y) :- \text{hasword}(X, \text{Word}), \text{hasword}(Y, \text{Word}).$ ” and rule “0.3:: $\text{sameauthor}(X, Y) :- \text{author}(X, \text{Title1}),$

*sametitle(Title1, Title2), author(Y, Title2).*” at the same time. We followed the Problog settings to use **noisy or** operation to estimate the joint probability:

$$P(X|R_1, \dots, R_k, \neg R_{k+1}, \dots, \neg R_n) = 1 - \prod_{i=1}^k (1 - p_i) \quad (1)$$

where  $p_i$  is the probability of rule  $R_i$  to be true. Noisy or is a probabilistic generalization of the logical or. In the previous example  $P(\text{sameauthor}(\text{person1}, \text{person2}) = \text{true}) = 1 - (1 - 0.2)(1 - 0.3) = 0.44$ .

#### 4.2. Structure and Parameter Learning

ILB can Learn Problog rule trees through many simple decision tree learners, for instance as C4.5 [Qui93] or CART [BFOS84]. In order to make these learners feasible, ILB will turn relational data into a feature-based discriminative dataset at first.

When learning a target predicate  $p$ , we define an instance  $x$  of ILB is a pair  $(P, y)$  consists of label  $y \in \{-1, 1\}$  and a conjunction of some grounded logical atoms  $P = t_1 \wedge \dots \wedge t_n$ . Notice the set of **generated instances**  $E$  is different from the set of pre-labeled training examples  $\mathcal{E}$  defined in section 3. The outline of instance generation procedure is presented in algorithm 1.

The procedure of feature-based data extraction is based on a hypergraph generated from the original data. Relational database can be viewed as a hypergraph with constants as nodes, and true ground atoms as hyperedges. Therefore, first-order rule which defines a goal concept can be viewed as a template subgraph consists of numerous variabilized hyperedges.

Logical induction is actually to find such a template subgraph that could match (or cover) as much as positive examples and as few as the negative examples. However, the hypothesis space of subgraph structure is so huge and complex that makes ordinary searching algorithms intractable.

ILB constrains the searching space by relational path finding [RM92]. It is based on the assumption that there usually exists a fixed-length path of relations linking the set of terms that satisfying the goal concept. This approach achieves many good results in inductive logic programming area and inspired lots of SRL structure learning algorithms [MM07, KD09]. ILB uses

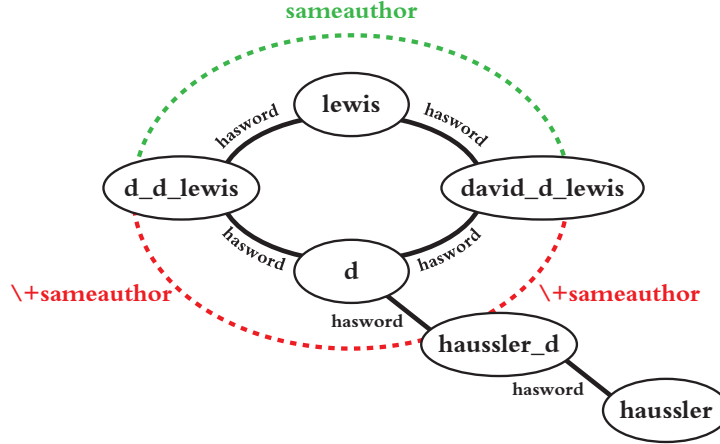


Figure 2: Example in Cora dataset of learning  $sameauthor(X, Y)$ : There are 3 people in this domain, their name have a common word “d”, but *haussler\_d* is different with the other nodes. “\+” means negation in Prolog clause.

paths of positive examples as the *core forms* from which it expands the searching space to find the optimal solution.

At the beginning of learning Problog Rule Trees, we construct a hypergraph  $\mathcal{G} = (E(\mathcal{G}), V(\mathcal{G}))$  from the relational database  $D$ , where  $E(\mathcal{G})$  is the set of hyperedges  $\{t_i(v_1, \dots, v_k), k = arity(t_i)\}$  (denoted by  $t$  because they are also *ground terms* in  $D$ ),  $V(\mathcal{G})$  is the set of all vertices in  $\mathcal{G}$ . This step is described as initialization step in algorithm 1.

Second, ILB uses a depth-first *FindPath* procedure to search paths that contains the nodes appear in a positive examples of  $e^+ = \varepsilon^+$ . In the example showed in figure 2, given a positive example  $sameauthor(d\_d\_lewis, david\_d\_lewis)$ , *PathFinding* will finds 2 grounded paths:  $P_1 = \{hasword(d\_d\_lewis, lewis), hasword(david\_d\_lewis, lewis)\}$  and  $P_2 = \{hasword(d\_d\_lewis, d), hasword(david\_d\_lewis, d)\}$ .

Third, variabilize the retrieved paths to get a **core form**  $h$ :  $-C$ , formally put as  $\theta e :- \theta P$ , where  $\theta = [a_i/A_i]$  is a substitution replaces each unique constant appear in  $P$  with a variable. In the above example we’ll get only one core form  $sameauthor(X, Y) :- \{hasword(X, Z), hasword(Y, Z)\}$ . This step is described as *Substitute* procedure in algorithm 1.

Fourth, to capture the underlying global distribution in domain, ILB uses the *core forms* to match the whole hypergraph to find all grounded paths that satisfy it, this step can be efficiently done by querying the first-order clause of the core form in the relational database, shown as *QueryProlog*



procedure in algorithm 1. In the example, we will get four grounded paths: the former two paths  $P_1$ ,  $P_2$  and two new paths  $P_3 = \{hasword(haussler\_d, d), hasword(d\_d\_lewis, d)\}$  and  $P_4 = \{hasword(haussler\_d, d), hasword(david\_d\_lewis, d)\}$  (In this example there will be another four symmetric paths, here we omit them as trivial situations).

Finally, we construct an instance  $x = (P, y)$  for training data as follows: for each retrieved path  $P$  check its deduced ground atom deduced from the core form, which is described as *PrologDeduce* in algorithm 1. If the grounding belongs to  $\varepsilon^+$  then assign  $y = 1$ , if it belongs to  $\varepsilon^-$  or does *not appears* in  $\varepsilon$  then  $y = -1$  (This follows the closed world assumption). Hence the algorithm can generate 4 instances in our example,  $P_1$  and  $P_2$  that deduce *sameauthor(d\\_d\\_lewis, david\\_d\\_lewis)* will be marked as positive instances;  $P_3$  and  $P_4$  deduce *sameauthor(haussler\\_d, david\\_d\\_lewis)* and *sameauthor(haussler\\_d, d\\_d\\_lewis)*, which are not in  $\varepsilon^+$ , so they will be marked as negative instances.

After we get the labeled instances, a more important task for rule induction by statistical learning is to calculate features. ILB uses two kinds of feature for each generated instance  $x = \{P, y\}$ :

- **Branch Feature:** A *branch feature*  $\Phi_b(P, v_i)$  of node  $v_i$  in path  $P$  is a substituted path  $\theta P'$ , in which  $P'$  starts from  $v_i$  and only share one node with  $P$ . Thus it looks like a tree branch spread out from  $P$ .  $\theta$  is the same substitution that maps  $P$  to its core form  $C(P)$ . In the example of figure 2, instance  $P_4$  has a branch feature *hasword(X, haussler)*.
- **Path Feature:** some vertices in path  $P$  might be also connected by paths other than  $P$ . These paths have more than one shared vertices with  $P$ , thus each one of them can form loops with some edges in  $P$ . We substitute them with  $\theta$  and variabilize all other unique constants to construct *path features*. Without loss of generality, we define *path feature*  $P' = \Phi_p(P)$  are the paths that their start node and end node are the only 2 shared nodes with  $P$ . If  $P'$  have  $n > 2$  shared nodes with  $P$ , then it can be split into  $n - 1$  parts that each part only shares 2 nodes with  $P$ . In the previous example, instance  $P_1$  has a path feature  $\{hasword(X, Z1), hasword(Y, Z1)\}$  where  $Z1=d$  is different from  $Z=lewis$  formulates another path from  $X$  to  $Y$ .

The *branch feature*  $\Phi_b(P, v_i)$  represents individual property of each node  $v_i$  in path  $P$ , the *path feature*  $\Phi_p(P)$  captures auxiliary relations of the nodes within  $P$ .

---

**Algorithm 1** *GenerateInstance*( $\mathcal{B}, \varepsilon^+$ )

---

**Input:** Background knowledge (Relational Database)  $\mathcal{B}$ , positive examples  $\varepsilon^+$

**Output:** Training data  $D = \{(\mathbf{P}_i, y_i)\}$

Initialize hypergraph  $\mathcal{G}$  with  $\mathcal{B}$ .

$D = \phi$ .

$coreForms = \phi$

**for** each example  $e \in \varepsilon^+$  **do**

$paths = FindPath(\mathcal{G}, e)$

**for** each path  $\mathbf{P} \in paths$  **do**

        core form  $c = Substitute(\mathbf{P})$

**if**  $C \notin coreForms$  **then**

            add  $C$  to  $coreForms$

**end if**

**end for**

**end for**

**for** each core form  $C \in coreForms$  **do**

$allPaths = QueryProlog(\mathcal{B}, c)$

**for** each queried output  $\mathbf{P}_i \in allPaths$  **do**

$head = PrologDeduce(\mathbf{P}_i, c)$

**if**  $head \in \varepsilon^+$  **then**

$y_i = 1$

**else**

$y_i = -1$

**end if**

$(\Phi_b(\mathbf{P}_i), \Phi_p(\mathbf{P}_i)) = computeFeature(\mathbf{P}_i)$

$x_i = generateInstance(c, \mathbf{P}_i, y_i, \Phi_b(\mathbf{P}_i), \Phi_p(\mathbf{P}_i))$

        add  $x_i$  to  $D$

**end for**

**end for**

**Return:**  $D$

---

With these two kinds of features, Inductive Logic Boosting can expand the core form to estimate the goal concept by greedily adding the best feature step by step.

**Theorem 1** (Completeness of Features) *If the optimal body of the goal predicate  $p(A_1, \dots, A_n)$  can be represented by a variabilized hypergraph  $\mathcal{G}_p$ , then starting a expansion from any core form  $C$  that contains all arguments  $A_i$  of  $p$ , we can find all other hyper edges belongs to the optimal  $\mathcal{G}_p$  only by exploring and variabilizing those paths in  $\Phi_b(\mathbf{P})$  and  $\Phi_p(\mathbf{P})$ , where  $\mathbf{P}$  is a grounded path of any positive instances  $x = (\mathbf{P}, 1)$  of predicate  $p$ .*

**Proof:** Notice that a positive instance  $\mathbf{P}$  is generated from a positive example  $p(a_1, \dots, a_n)$  which unifies with the goal predicate  $p(A_1, \dots, A_n)$ , there exists a core form  $C \in \mathcal{G}_p$  whose body can be grounded to  $\mathbf{P}$ . Moreover, since  $\mathbf{P}$  is positive, there also exists a grounded graph  $g_p$  that contains  $\mathbf{P}$  and unifies with the optimal first-order graph  $\mathcal{G}_p$ . Because the number of vertices in  $\mathcal{G}_p$  is finite, the optimal substitution  $\theta^*$  from node to variables can be easily found. Therefore, the goal of this proof has been reduced to prove the grounded graph  $g_p$  is reachable when expanding  $x$  with only  $\Phi_b(\mathbf{P})$  and  $\Phi_p(\mathbf{P})$ .

Because  $\mathcal{G}_p$  is connected, so after substitution  $\theta^*$  the grounded graph  $g_p$  remains connected. Hence for each hyper edge  $t' \in g_p \wedge t' \notin \mathbf{P}$ , there must exists at least one path  $\mathbf{P}'$  that starts from  $t'$  and ends with a node  $v \in \mathbf{P}$ . So for all hyper edges  $t' \in g_p$ , there exists  $\mathbf{P}' \in \Phi_b(\mathbf{P})$  that contains  $t'$ . For the hyper edges  $t'$  who can connect  $\mathbf{P}$  through more than one paths in  $\Phi_c(x)$ , e.g.  $\mathbf{P}'_1$  and  $\mathbf{P}'_2$ , we can construct a *path feature* by connecting  $\mathbf{P}'_1, \mathbf{P}'_2$ .  $\square$

We can see that with only  $\Phi_b$  is already complete for searching the optimal rule of goal concept. However, the number of  $\Phi_b$  increases exponentially with the length of path, which means the information conveyed by *branch feature* decreases exponentially by its length. Conversely, *path features* is more informative. They provide auxiliary relation information. Like the example in figure 2, to learn a predicate  $sameauthor(X, Y)$  start from core form  $\{hasword(X, Z), hasword(Y, Z)\}$ , apparently a path feature  $\{hasword(X, Z1), hasword(Y, Z1)\}$  as more important than branch features like  $\{hasword(X, Z2)\}$ .

The *computeFeature* procedure in algorithm 1 represents this step. Finally with all paths and features we computed, the training instances are generated and added to training data. For the

example in figure 2, a part of finally generated data for learning predicate  $sameauthor(X, Y)$  is displayed in table 1.

Core Form	$sameauthor(X, Y):-hasword(X, Z), hasword(Y, Z).$	
$x_1$	$y$	1
	P	$\{hasword(d\_d\_lewis, lewis), hasword(david\_d\_lewis, lewis)\}.$
	$\Phi_b$	$\{hasword(X, d)\}.$ $\{hasword(Y, d)\}.$
	$\Phi_p$	$\{hasword(X, Z1), hasword(Y, Z1)\}.$
$x_2$	$y$	-1
	P	$\{hasword(haussler\_d, d), hasword(david\_d\_lewis, d)\}.$
	$\Phi_b$	$\{hasword(X, haussler)\}.$ $\{hasword(Y, lewis)\}.$ $\{hasword(Y, david)\}.$
	$\Phi_p$	—

Table 1: Example of generated instance in Cora dataset

From the generated labeled data, we can learn a Problog Rule Tree for each *core form* by any decision tree learner.

#### 4.3. Boosting

Inductive Logic Boosting learns a discriminative task rather than doing logic induction by calculating coverage. This feature makes ILB more fits to boosting framework.

ILB use the confidence-based Adaboost [SS99] in boosting stage. However, the weak hypothesis  $h_t$  of each round boosting is a set of Problog rules, which are expanded from different core forms. To combine them, ILB use the *noisy or* feature we have mentioned before. During each evaluation of current hypothesis  $h_t$ , for all instances  $x \in Instance_e = \{x | PrologDeduce(x, C) = e\}$  that can deduce same example  $e \in \varepsilon$  with its core form, ILB assigns the probability  $P(x) = NoisyOr(h_t(x_1), \dots, h_t(x_m)), \forall x_i \in Instance_e$ . Notice that there might be some instances  $x_k = (P_k, y_k)$  that  $h_t$  not covered, follows the *closed world assumption*, we define  $h_t(k) = 0$ .

Dataset	Types	Constants	Predicates	True Atoms	Total Atoms
Cora	5	3,079	10	42,558	687,422
UW-CSE	9	929	12	2112	260,254

Table 2: Detail of datasets

## 5. Experiments

### 5.1. Datasets

We carried out experiments on two real world datasets to investigate whether ILB performs better than previous state-of-the-arts SRL approaches. The task is to learn target predicate definitions with evidence predicates. Both datasets are publicly available at <http://alchemy.cs.washington.edu>.

**Cora.** This dataset is a collection of citations to computer science papers, created by Andrew McCallum, and later processed by Singla and Domingos [PD07] into 5 folds for the task of duplicating the citations. Evidence predicates are other relations like *author(Bib, Author)*, *title(Bib, Title)*, *venue(Bib, Venue)*, and so on. Relations in this domain is simple and clear.

**UW-CSE.** This dataset was prepared by Richardson and Domingos [RD06], describes relationships in an academic department. The dataset is divided into 5 independent areas/folds (AI, graphics, etc.). The evidence predicates describe students, faculty, and their relationships (e.g, *Professor(person)*, *TaughtBy(course, person, quarter)*, etc.). Target predicate is *advisedBy(Person, Person)*. We omitted 9 equality predicates follows [KD10]. Relational structure of this domain is more complex since predicate and arguments are more complicate than Cora domain.

The detail of each dataset is showed in 5.1. Cora has more constants but has a simpler and clearer relation structure, UW-CSE is a more complex relational model (hyper graph).

### 5.2. Compared Methods

We compared Inductive Logic Boosting to following state-of-the-art systems:

**RDN-Boost** [NKK<sup>+</sup>12]. This algorithm represents a Relational Dependency Network (RDN) model as regression trees and learns by boosting. It turns the structure learning problem into

a series of relational function-approximation problems and solves by gradient-boosting, which easily induces highly complex features over several iterations and in turn estimate quickly a very expressive model. This work outperforms numbers of state-of-the-art SRL structure learning algorithms. Base on this system there is also a modified version for learning Markov Logic Network [KNKS11]. This approach is denoted as RDN-B.

**Learning MLN structure by Structure Motif** [KD10]. Key insight of this approach is that relational data usually contains recurring patterns, which is called structural motifs. By constraining the search for clauses to occur within motifs, it can greatly speed up the search and thereby reduce the cost of finding long clauses(i.e., formulas with more than 4 or 5 literals). We use LSM to denote it.

In order to make the comparison as fair as possible, we used the following protocol. For RDN Boost, we use the default parameter setting that constrain maximum tree hierarchy to be 4, each node contains 2 literals at most and boosting for 20 turns. For LSM, we employ the parameter suggested in [KD10]. ILB searches uses same hierarchy and node literal length (in ILB we constrain the path length in feature  $\Phi_b$  and  $\Phi_p$ ) limit settings as RDN Boost since they both learn clauses based on boosting weak learners in tree structure. Besides, we constrained the max *core form* length for Cora and UW-CSE to be 4 and 2 accordingly.

Notice that both RDN and LSM have to input background knowledge of constant types and predicate forms (e.g. *author(Bib, Author)* indicates predicate *author* only can take *Bib* and *Author* as arguments in exact those position). Further more, for RDN approaches we enumerated all possible predicate “modes” (e.g. *samebib('Bib, +Bib)* indicates the first *Bib* can be not in the head of learned clause) to ensure the completeness while learning clauses. ILB does not need to use predefined predicate formulation or variable types. However, the instance generation procedure always produces huge number of instances (especially negative instances). Therefore, we randomly sample a part of them during instance generation. In Cora task, we randomly generate 1000 instances for each *core form* in 1 fold, and 300 in UW task.

### 5.3. Results

We evaluate these approaches not only by the labeled positive and negative examples. Actually, there always be many falsities for a target concept that not covered by the labeled examples. Thus

	SameAuthor		SameBib	
System	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC
ILB	<b><math>0.9517 \pm 0.05</math></b>	<b><math>0.9835 \pm 0.02</math></b>	<b><math>0.9576 \pm 0.04</math></b>	<b><math>0.9967 \pm 0.00</math></b>
RDN-B	$0.8094 \pm 0.14$	$0.8877 \pm 0.13$	$0.9046 \pm 0.03$	$0.9475 \pm 0.02$
LSM	–	–	–	–
	SameTitle		SameVenue	
System	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC
ILB	<b><math>0.7668 \pm 0.09</math></b>	<b><math>0.9784 \pm 0.02</math></b>	<b><math>0.6696 \pm 0.11</math></b>	<b><math>0.9606 \pm 0.01</math></b>
RDN-B	$0.1424 \pm 0.05$	$0.7790 \pm 0.06$	$0.0855 \pm 0.03$	$0.5698 \pm 0.03$
LSM	–	–	–	–

Table 3: Results on Cora dataset

	advisedBy	
System	AUC-PR	AUC-ROC
ILB	$0.6192 \pm 0.16$	$0.8932 \pm 0.08$
RDN-B	$0.6140 \pm 0.21$	$0.9549 \pm 0.03$
LSM	$0.22 \pm 0.02$	$0.62 \pm 0.06$

Table 4: Results on UW-CSE dataset

we evaluate the predicate concept induction task in global distribution by treating all relations that not appears in labeled data as negative examples.

Consider the learned definition of target predicate as a binary classifier, we choose AUC-ROC to compare these approaches. However, a key property of most relational data sets is the number of negatives can be order of magnitude more than the number of positives. To ignore the impact of the overwhelming true negatives, we also use area under precision-recall curves (AUC-PR) to evaluate the performance.

The evaluation result of Cora dataset is showed in table 3. In this task LSM only produces trivial unit clauses, so ILB is only compared with RDN-Boost. We can see that both AUC-PR and AUC-ROC value of ILB are significantly better than RDN-Boost. A major reason is that the RDN-B learns a graphical model by maximizing pseudo-likelihood which assumes independence between all random variables (logical literal) in logic formulas, while ILB directly uses the empirical

probability to estimate possibility of a rule being satisfied, which results in better estimation of the goal concepts.

Result on UW-CSE is presented in table 4. In facts, due to the high complexity in hypergraph generated in by UW-CSE, path finding in the relational graph will get so many paths, which results in more than 100 core forms and millions of feature-based instances and path features, which hardly can be handled by ILB. Thus, we only samples 5% of instances and 10% of features to do Problog Tree induction. With a highly incomplete training data, the result is still comparable with RDN-Boost. LSM performs worst because we did not run another round weight learning procedure as [KD10] suggests. But even if we do, the performance of LSM should also be worse than RDN-Boost [KNKS11].

We can also observe that the results in table 3 for RDN-Boost and LSM are worse than those reported by [NKK<sup>+</sup>12], [KNKS11] and [KD10]. They learn a predicate with all other predicates as evidences. For instance, when learning *SameBib* in Cora, *SameAuthor* and *SameTitle* can be used in body of learned hypothesis. Our task on predicate concept induction is much more challenging since we do not use any predicates that can be superseded by other predicates in the domain. This ability ensures us to discover novel knowledge from the most basic concepts in a domain.

## 6. Conclusions

We presented Inductive Logic Boosting, which learns weighted logical rules in a statistical learning framework. It uses path-finding in relational domain to generate a discriminative labeled dataset and calculates two kinds of features. Then performs decision tree boosting, a simple yet effective statistical learning algorithm, to learn Problog rules. Our empirical comparisons with two state-of-the-art systems on real datasets demonstrate the effectiveness of ILB.

As future work, we want to generalize the ILB framework to accomplish more logic induction tasks like predicate invention and learn recursive rules, to provide a different angle of view for inductive logic programming.



## References

- [BFOS84] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [BR13] Elena Bellodi and Fabrizio Riguzzi. Structure learning of probabilistic logic programs by searching the clause space. *Computing Research Repository*, abs/1309.2080, 2013.
- [DBR13] Nicola Di Mauro, Elena Bellodi, and Fabrizio Riguzzi. Bandit-based monte-carlo structure learning of probabilistic logic programs. In *23rd International Conference on Inductive Logic Programming*, pages 1093–1097, Washington, DC, 2013. IEEE.
- [DRK08] Luc De Raedt and Kristian Kersting. *Probabilistic Inductive Logic Programming*. Springer, 2008.
- [DRKT07] Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. Problog: A probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2462–2467, Hyderabad, India, 2007.
- [GT07] Lise Getoor and Ben Taskar. *Introduction to statistical relational learning*. The MIT press, 2007.
- [KD05] Stanley Kok and Pedro Domingos. Learning the structure of markov logic networks. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 441–448, Bonn, Germany, 2005. ACM.
- [KD09] Stanley Kok and Pedro Domingos. Learning markov logic network structure via hypergraph lifting. In *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, 2009. ACM.
- [KD10] Stanley Kok and Pedro Domingos. Learning markov logic networks using structural motifs. In *Proceedings of the 27th International Conference on Machine Learning*, pages 551–558, Haifa, Israel, 2010. Omnipress.
- [KNKS11] Tushar Khot, Sriraam Natarajan, Kristian Kersting, and Jude W. Shavlik. Learning markov logic networks via functional gradient boosting. In *11th IEEE International Conference on Data Mining*, pages 320–329, Vancouver, BC, Canada, 2011. IEEE.

- [LD94] Nada Lavrac and Saso Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, 1994.
- [MM07] Lilyana Mihalkova and Raymond J. Mooney. Bottom-up learning of markov logic network structure. In *Proceedings of 24th International Conference on Machine Learning*, Corvallis, OR, 2007.
- [Mug91] Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, 1991.
- [NJFH03] Jennifer Neville, David Jensen, Lisa Friedland, and Michael Hay. Learning relational probability trees. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, Washington, DC, 2003. ACM.
- [NKK<sup>+</sup>12] Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Gutmann, and Jude W. Shavlik. Gradient-based boosting for statistical relational learning: The relational dependency network case. *Machine Learning*, 86(1):25–56, 2012.
- [PD07] Hoifung Poon and Pedro Domingos. Joint inference in information extraction. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 913–918, Vancouver, British Columbia, Canada, 2007. AAAI Press.
- [Qui93] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA, 1993.
- [Qui96] J. Ross Quinlan. Boosting first-order learning. In *Proceedings of 7th International Workshop on Algorithmic Learning Theory*, pages 143–155, Sydney, Australia, 1996. Springer Berlin Heidelberg.
- [RD06] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.
- [RM92] Bradley L. Richards and Raymond J. Mooney. Learning relations by pathfinding. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 50–55, San Jose, CA, 1992. AAAI Press.

- [SS99] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.